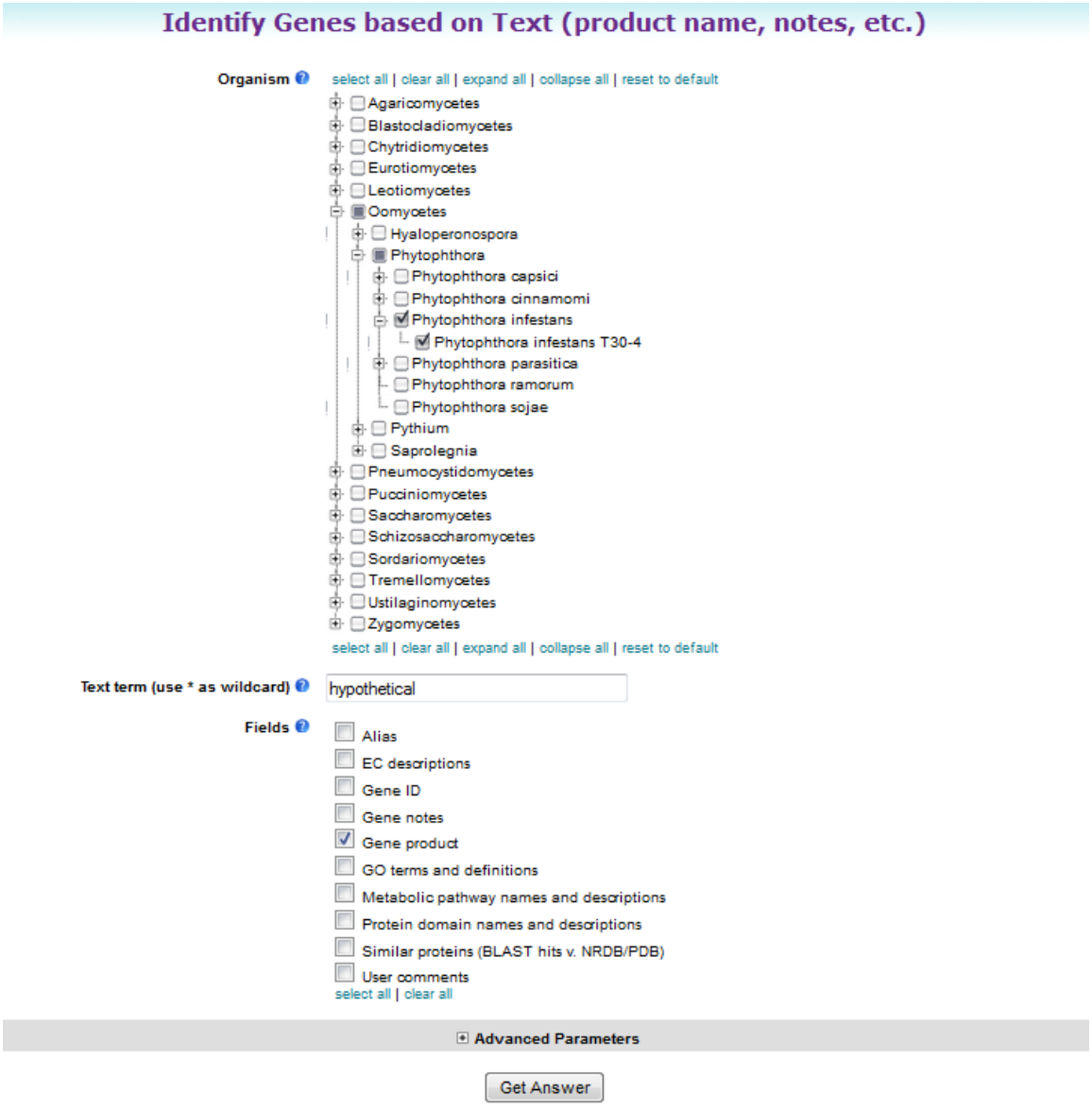
Protein Motif Searches and Regular Expressions
Exercise 6

6.1 Using InterPro domain searches to identify unannotated kinesin motor proteins.

For this exercise use http://fungidb.org

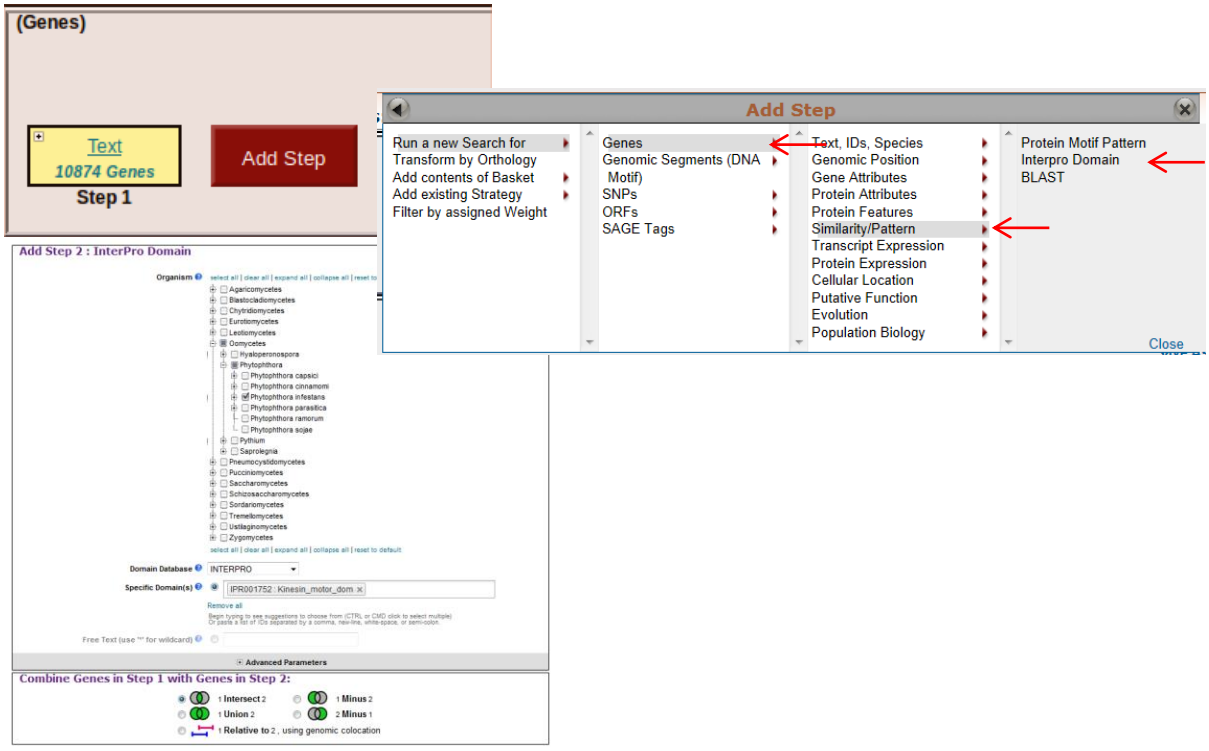
a. Identify all genes annotated as hypothetical in *Phytophthora infestans.*

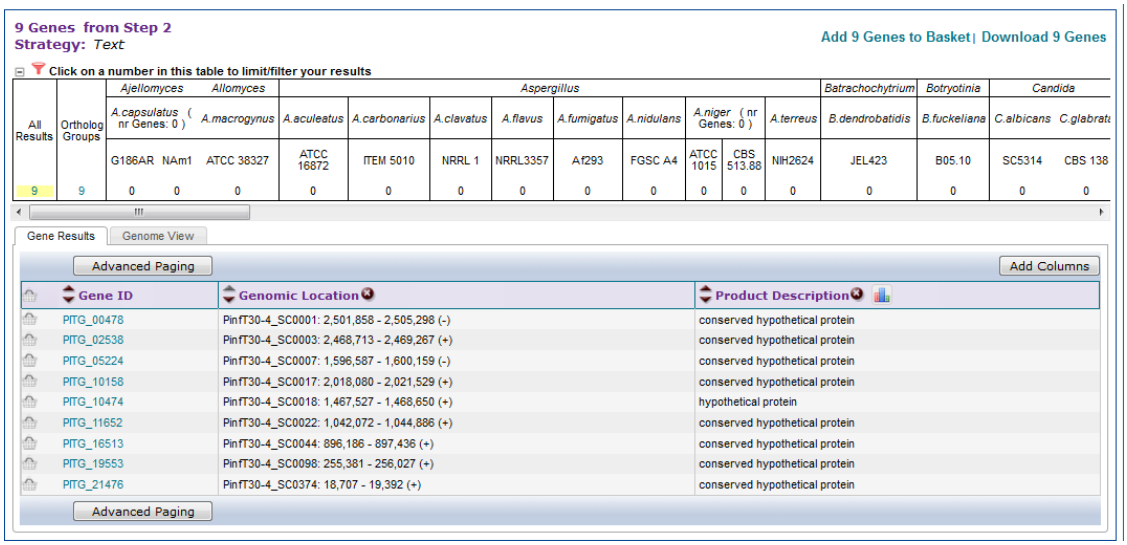
Hint: use the full text search and look for genes with the word “hypothetical” in their product names.



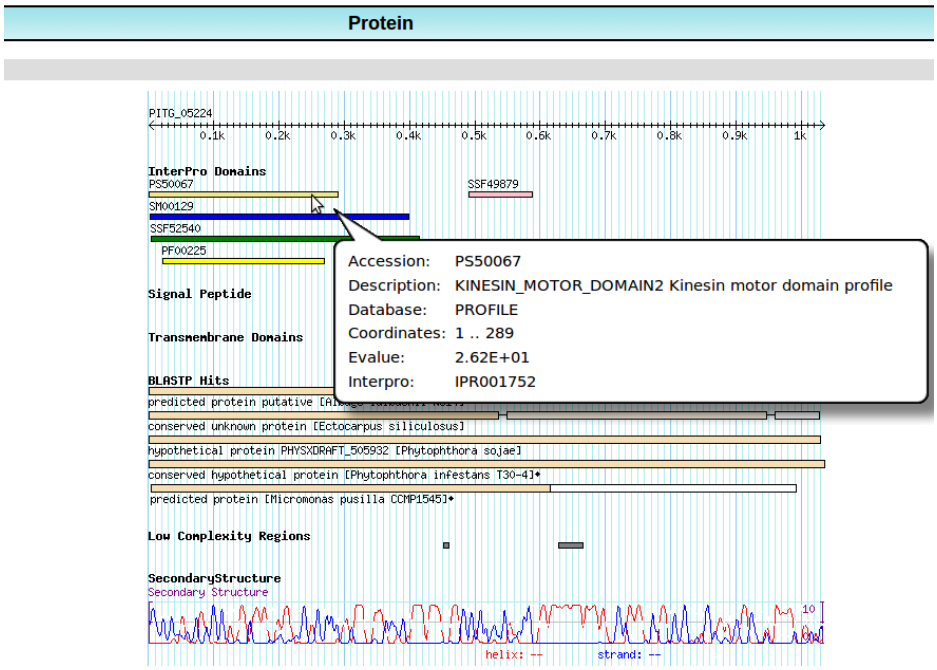
b. How many of these hypothetical genes have a kinesin-motor protein InterPro domain?

Hint: add a step to the strategy. Go to the “Interpro Domain” search under similarity/pattern, start typing the work kinesin and it should autocomplete.



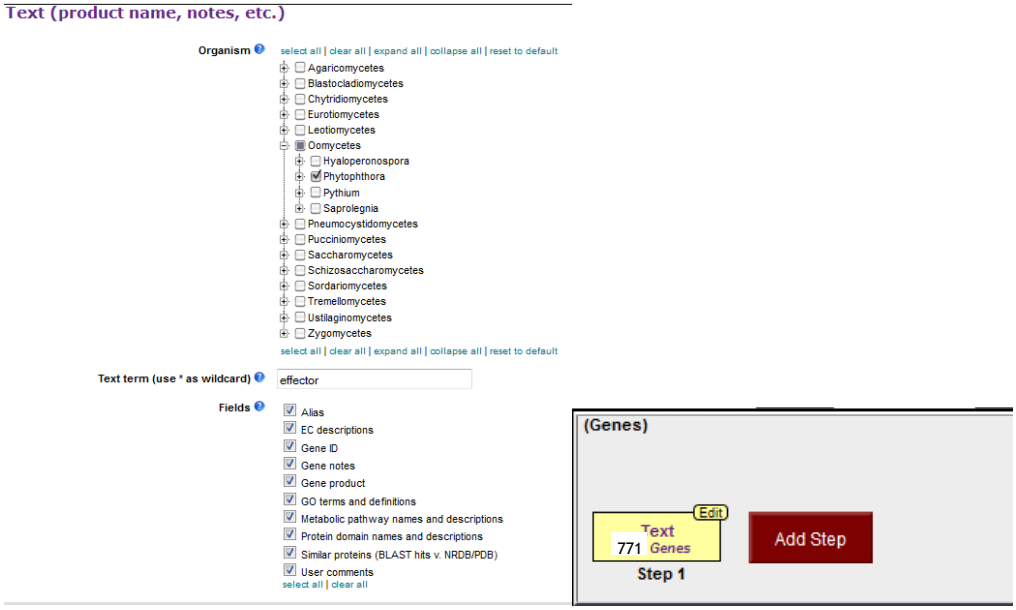


c. Go to the gene page for PITG_05224 and look at the protein feature section. Does this look like a possible motor protein?
Hint: click on the ID for PITG_05224 in the result table to go to the gene page. Mouse over the glyphs in the Protein Features graphic.

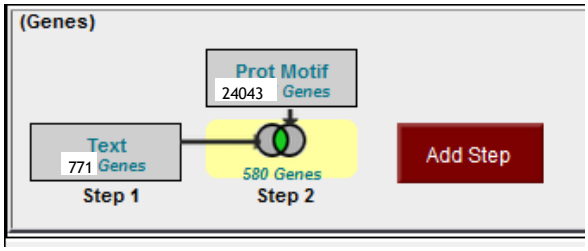


6.2 Using regular expressions to find motifs in Phytophthora. Find variations of RXLR

a. To infect plants Phytophthora utilizes effector proteins. Use a text search to find all proteins that have been identified as effectors in Phytophthora.



b. RXLR is a domain motif found in some effectors to facilitate infection. Identify all occurrences of the RXLR motif in Phytophthora. You may need to refer to the RegEx guides to find the correct query; you will need to use a special character for 'X'.

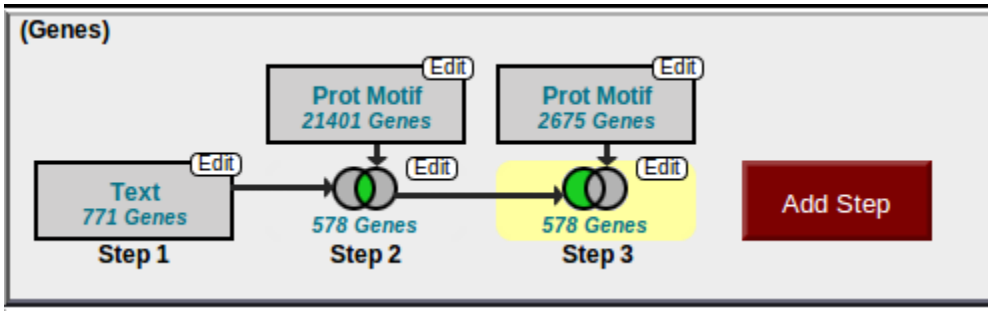


c. Some of these were probably identified in incomplete proteins. You could use a text search to omit predicted or hypothetical proteins, but instead think of a protein motif search to only identify complete proteins containing RxLR. Protein sequences in FungiDB do not contain the stop character (*). However, bad computationally predicted proteins can have internal stops. Edit your motif search to select for proteins that start with a Methionine, do not have any *s, and contain an RXLR.

   (hint: you'll need to tell the RegEx to not find * both before and after the RXLR)

   You can find a single RegEx to identify the correct proteins but it will be complex. Try to break it up into multiple steps to make it easier to build.

   Here it is split into two RegEx:



   It only removed two genes. Why? Compare the results from B and C, where was the change?

d. The 'X' in RXLR is a wild-card, allowing for any amino acid. Try some specific amino acids or special characters to narrow down the RXLR occurrences. Do most identified RXLRs fit into any special classification?